Использование интерактивной визуализации в задаче извлечения признаков из слабоструктурированных текстовых данных

Е.А. Макарова [1], Д.Г. Лагерев [2]

[1] ООО «К-Скай», наб. Варкауса, 17, Петрозаводск, Респ. Карелия, 185031, Россия
[2] Брянский государственный технический университет, бул. 50 лет Октября, 7, Брянск, 241035, Россия

Аннотация

В статье рассматриваются визуализации слабоструктурированных текстовых данных (ССТД) с целью решения задач разведочного анализа и построения модели обработки текстовых данных для их дальнейшего использования в моделях анализа данных. Рассмотрены проблемы, с которыми сталкиваются исследователи при добавлении ССТД в модель анализа данных. Рассмотрены существующие подходы к визуализации текстовых данных для решения различных задач обработки естественного языка. Предложена модель интеллектуальной обработки ССТД и подходы к трансформации данных в процессе обработки. Для визуализации процесса трансформации ССТД применяется визуальная модель, использующая диаграммы Санкей. Предложенная визуальная модель позволяет сократить время эксперта на обработку данных, благодаря повышению наглядности процесса извлечения признаков из ССТД и использованию интерактивных визуальных инструментов. Разработанный подход апробирован на данных, полученных из информационной системы службы занятости населения.

Ключевые слова

Обработка текстовых данных, разведочный анализ, визуализация, диаграмма Санкей.

Using Interactive Visualization in the Problem of Feature Extraction from Semi-structured Text Data

E.A. Makarova [1], D.G. Lagerev [2]

[1] K-SkAI Ltd., Varkaus qy., 17, Petrozavodsk, Rep. of Karelia, 185031, Russia
[2] Bryansk State Technical University, 50 let Oktyabrya blvd., 7, Bryansk, 241035, Russia

Abstract

The article deals with the visualization of semi-structured text data (SSTD) in order to solve the problems of exploratory analysis and build a model for processing text data for their further use in data analysis models. The problems faced by researchers when adding SSTS to the data analysis model are considered. Existing approaches to visualization of text data for solving various problems of natural language processing are considered. A model of intelligent processing of SSTD and approaches to data transformation within the data processing. A visual model used to visualize the process of transformation of SSTD is based on the Sankey charts. The proposed visual model allows to reduce the expert's time for data processing by increasing the visibility of the process of extracting features from SSTD using interactive visual tools. The developed approach was tested on data from the information system of the employment service.

Keywords

Text data processing, exploratory analysis, visualization, Sankey chart.

ГрафиКон 2023: 33-я Международная конференция по компьютерной графике и машинному зрению, 19-21 сентября 2023 г., Институт проблем управления им. В.А. Трапезникова Российской академии наук, г. Москва, Россия
EMAIL: m4karova.e@yandex.ru (Е.А. Макарова); lagerevdg@yandex.ru (Д.Г. Лагерев)
ORCID: 0000-0002-5410-5890 (Е.А. Макарова); 0000-0002-2702-6492 (Д.Г. Лагерев)

1. Введение

В информационных системах, поддерживающих ручной ввод информации различными людьми, обычно хранится множество текстовых данных различной структуры. Примерами подобных систем являются социальные сети, новостные порталы, сервисы трудоустройства, медицинские информационные системы и т.д. При построении моделей анализа данных с использованием информации, накопленной в подобных системах, целесообразно проверить на значимость как можно больше признаков, чтобы впоследствии учесть все необходимые свойства описываемых данными объектов и явлений. Сложность автоматического анализа подобных данных состоит в том, что большая часть из них является слабоструктурированными. Анализ слабоструктурированных текстовых данных (далее – ССТД) требует настройки системы сбора и предварительной обработки данных, которая будет учитывать и структуру обрабатываемых данных, и специфику решаемой задачи. Из ССТД могут быть извлечены ценные признаки, такие как: упоминание именованных сущностей, терминов, тем и пр. Исследователи отмечают разнообразие и сложность унификации методов предварительной обработки и подготовки ССТД к дальнейшему анализу [1], что приводит к существенному увеличению времени на подготовку данных при включении в них ССТД. Кроме того, при решении ряда задач для корректной обработки данных необходимо дополнительно привлекать специалиста в конкретной предметной области. Визуализация данных – один из способов упростить и ускорить процесс анализа и обработки ССТД.

Для компактного и эффективного представления больших массивов текстовых данных активно используются различные методы визуализации. Визуализация широко используется для представления данных разной структуры и сложности – от простых графиков, отражающих изменение одной переменной, до сложных специализированных представлений многомерных данных. Визуальная аналитика применяется на различных этапах исследования данных, включая этап разведочного анализа, который используется для построения гипотез и отбора данных с целью использования в моделях анализа данных [2]. Однако, большая часть визуальных моделей ориентирована на представление текущего состояния данных, в то время как для настройки обработки ССТД важно видеть процесс их трансформации. Разработка визуальных моделей, отражающих процесс преобразования данных на различных этапах их обработки, является целью данного исследования.

2. Визуализация слабоструктурированных текстовых данных в процессе обработки

2.1. Обзор существующих подходов к визуализации текстовых данных

Визуализация текстовых данных – развивающаяся область визуализации информации и визуальной аналитики. Разработано множество подходов и методов для решения широкого круга аналитических задач, в том числе визуализации, что позволяет исследователям из разных дисциплин извлекать информацию из ССТД. В качестве распространенных методов для визуализации текстовых данных, включая слабоструктурированные, можно перечислить: облака слов (тегов), UMAP [3], диаграммы уклона [4], диаграмма Санкей [5] и т.д. Активно развиваются методы, ориентированные на визуализацию моделей векторного представления слов [6], для реализации которых важно решить задачу снижения размерности.

Исходя из обзоров различных методов визуализации, не существует универсального метода, подходящего под все задачи обработки текстовых данных, из-за таких их свойств, присущих естественному языку, как высокая размерность, нерегулярность и неопределенность [7]. Среди существующих методов визуализации текстовых данных есть методы, работающие как с отдельными текстами, так и с коллекциями документов и корпусами текстов.

Выбор метода визуализации также зависит от задачи обработки естественного языка, как, например, тематическое моделирование, анализ тональностей, извлечение сущностей и т.д. Для

каждой из перечисленных задач используются различные методы визуализации. Визуализация возможна на различных этапах анализа текстовых данных, от разведочного анализа – до формирования отчёта о результатах анализа, который будет передан лицу, принимающему решение. Существуют исследовательские проекты, целью которых является сбор и систематизация визуализаций, применимых к текстовым данным [8].

В предыдущих публикациях авторами предлагались различные визуальные модели, построенные на основе визуализации «облако слов» [9], которая широко используется в различных задачах визуализации текстов. В данной работе рассматриваются визуальные модели текстовых данных на этапе разведочного анализа и настройки процесса обработки текстовых данных с целью решения задачи выделения ценных признаков из ССТД.

2.2. Интерактивная визуальная модель для извлечения признаков из слабоструктурированных текстовых данных

Распространенным способом использования текстовых данных в моделях анализа данных является выделение тем, упоминаемых терминов, именованных сущностей и использование их в качестве логических, категориальных или числовых переменных. Однако, если речь идёт о больших объемах ССТД, то одной из первоочередных задач является сокращение пространства используемых переменных до того объема, который возможно эффективно использовать как в моделях анализа данных, так и в визуальной аналитике так, чтобы это было доступно для исследования человеком. Кроме того, набор признаков лучше формировать совместно со специалистом в предметной области, чтобы отделить более важные переменные от менее важных, и корректно настроить процесс их извлечения.

Процесс формирования ограниченного набора признаков из большого количества извлеченных n-gram – сочетаний токенов (слов, чисел, символов), полученных из текста в процессе токенизации – это процесс объединения близких по значению n-gram в одну группу. Данное объединение может происходить, исходя из синтаксической и семантической близости. Так как мера близости может быть выбрана различная, в зависимости от исходных данных и решаемой задачи, данный процесс необходимо настраивать для каждого источника данных отдельно. Однако, в случае регулярно обновляющихся текстовых данных одной структуры, параметры обработки могут повторно использоваться для обработки новых данных.

В предыдущих работах авторов [10] рассматривалась модель интеллектуальной обработки ССТД, позволяющая настроить процесс обработки данных, исходя из их свойств, и улучшить качество ССТД для дальнейшего использования в процессе анализа данных. Для этого на первом шаге определяются параметры, свойственные ССТД из определенного источника. Одной из распространенных проблем текстовых данных, вводимых вручную, является наличие сокращений и орфографических ошибок, различная капитализация, наличие спецсимволов, использование различных обозначений для одного термина. Из-за этого одно и то же свойство описываемого текстом объекта или явления обозначается с помощью различных n-gram. При построении систем анализа данных их необходимо обозначать одним признаком, иначе падает точность анализа.

Параметры ССТД (P) в модели интеллектуальной обработки ССТД формально описываются следующим образом:

$$P = <L, E, D>,$$

где L – размер одного экземпляра данных;

E – процент ошибок и специфических сокращений, не входящих в словари общеупотребительных слов русского языка;

D – процент пар, состоящих из отдельных экземпляров данных со значением семантической близости выше заданного порога (по умолчанию 0,5).

Определить наличие подобных слов можно с помощью сравнения полученного при токенизации и лемматизации словаря и словаря общеупотребительной или профессиональной лексики:

$$E = \frac{Tokens \notin Dict}{Tokens} \cdot 100\,,$$

где *Tokens* – массив токенов, приведенных к лемме (нормальной форме), полученных в результате обработки ССТД;
Dict – список слов (словарь), который может состоять как из общеупотребительных слов используемого языка, так и включать профессиональную лексику.

Существуют различные способы раскрытия сокращений или исправления ошибочных написаний, в том числе те, что были предложены в предыдущих работах авторов [11]. Независимо от выбранного алгоритма, эксперту, работающему с данными, необходимо будет оценить результат обработки. Это можно сделать как с помощью выборочного анализа обработанных текстов, так и оценив изменение количества уникальных n-gram после преобразования и отдельные преобразования. Эти сведения также можно получить с помощью визуализаций данных, отражающих различные их свойства – например, использовать диаграммы, отражающие изменения числовых значений уникальных терминов.

Параметр *D* определяется следующим образом:

$$D = \frac{\sum_n^i J(A_i, B_i) < k''}{S} \cdot 100,$$

где *J* – выбранный коэффициент семантической близости;
k'' – коэффициент семантической близости, используемый для предварительной оценки степени дублирования в документах, по умолчанию 0,5,
n – количество текстовых документов в случайно выбранном для оценки дублированности наборе данных,
S – количество пар текстовых документов в случайно выбранном для оценки степени дублированности наборе данных,
A_i, B_i – экземпляры данных, семантическая близость между которыми измеряется.

Таким образом, стандартный конвейер может состоять из следующих этапов:

1) предварительная очистка и предобработка;
2) исправление ошибок;
3) раскрытие сокращений;
4) объединение n-gram на основе различных семантических отношений (синонимия, гипонимия, ассоциативность), рассчитываемых по уровню семантической близости.

В предварительную очистку и предобработку могут входить следующие этапы:

1) приведение к одному регистру;
2) удаление спецсимволов (%,+ и т.д.);
3) удаление чисел;
4) удаление стоп-слов.

В предыдущих работах было доказано [11], что порядок выполнения этапов по предварительной обработке данных влияет на качество полученных данных, и что изменения ССТД, произведенные на предыдущих этапах, влияют на результат выполнения последующих.

Таким образом, для такого многоступенчатого и контекстно-зависимого процесса как обработка ССТД, важно видеть наиболее полную картину изменения данных при прохождении всех преобразований. Для того, чтобы проследить путь преда отдельных n-gram, предлагается использовать интерактивную визуальную модель, построенную на основе диаграммы Санкей.

После расчета основных параметров текстовых данных по формулам, представленным ранее, и анализа содержания текстовых документов на предмет наличия чисел и спецсимволов, выстраивается «рекомендуемый» конвейер обработки данных. Далее с помощью визуальных инструментов, таких как диаграммы, отображающие динамику уникальных n-gram, и диаграммы Санкей, исследователь сможет не только оценить процесс преобразования и изменить общие настройки обработки, но и отредактировать объединение/разъединение n-gram.

Так как при анализе больших массивов ССТД, даже после процесса объединения в группы, потенциальных признаков будет выделено много, чтобы эффективно учитывать их все, необходимо выделить признаки, которые будут наиболее ценны для решения задачи.

Определить список признаков может эксперт в предметной области, опираясь на свои знания и опыт. Существуют и автоматические методы отбора признаков для моделей анализа данных, позволяющие оценить «вес» отдельных признаков и их влияние на другие переменные в исследуемых данных. В качестве метрик для оценки «веса» переменных в моделях анализа данных могут использоваться различные коэффициенты корреляции, WOE index и т.д. [12] Эффективным способом определить оптимальный набор признаков, извлекаемых из ССТД, будет объединение человеческой экспертизы и предварительных расчетов «веса» отдельных признаков, полученных в процессе разведочного анализа.

При работе с интерактивной визуальной моделью эксперт должен минимизировать ошибки первого и второго рода, которые могут возникнуть в процессе извлечения признаков из ССТД, а именно:

1) пропуск признака в тексте, где он фактически присутствует;
2) извлечение признака из текста, где сам признак отсутствует, но имеются слова или словосочетания, чья семантическая близость с одним из вариантов записи признака выше выбранного порога.

3. Использование интерактивной визуальной модели на примере извлечения признаков из текстовых полей резюме соискателей

Для апробации предложенного подхода были использованы открытые данные из обезличенных резюме, опубликованных на сайте «Работа России» [13]. Анализ данных о поведении участников рынка труда может основываться на многих показателях как, например, соотношение соискателей и вакансий по определенным профессиональным отраслям и регионам, количество откликов и приглашений, и т.д. В резюме соискателей содержатся как структурированные данные – уровень образования, годы работы на определенной должности – так и слабоструктурированные текстовые данные в описании навыков, личностных качеств, достижений, полученных во время учебы или на предыдущих местах работы. Вся эта информация имеет влияние на принятие работодателем решения о приглашении соискателя на собеседование. Анализ влияния упоминания тех или иных свойств резюме соискателей на количество приглашений на собеседования может помочь в рамках реализации государственных программ содействия занятости при определении перспективных программ для профессиональной переподготовки временно безработных, а также результаты анализа могут быть полезны соискателям при составлении резюме.

Для настройки модели интеллектуальной обработки ССТД из резюме соискателей было выбрано 60 000 резюме, без фильтрации по региону или профессии, чтобы получить наиболее репрезентативную выборку для настройки универсального конвейера обработки резюме. Первым шагом работы модели является оценка различных параметров обрабатываемых данных, исходя из которых будет выстроен конвейер для обработки. Вторым шагом работы модели интеллектуальной обработки является создание модели Word2Vec на данных из предметной области, чтобы далее использовать модель векторного представления слов для определения семантической близости как отдельных токенов, так и словосочетаний, фраз.

Оценка параметров показывает, что в изучаемых данных встречаются несловарные (ошибочные или узкоспециализированные) слова в большом количестве. Анализ «соседей» некоторых токенов в модели векторного представления слов показывает, какие варианты написания термина встречаются в выборке. Например, ошибочные написания названия одного из распространенных программных продуктов, представлены в таблице 1.

Этап исправления ошибок, который включается в модель обработки автоматически при превышении определенного порога (по умолчанию это более 1% слов), может быть пропущен, если необходимо учитывать наличие орфографических ошибок для анализа резюме соискателя.

После построения стандартного конвейера обработки и подготовки ССТД, эксперт может оценить, насколько каждый этап повлиял на количество уникальных записей в целом, используя

табличные данные или визуализацию на основе столбчатых диаграмм (рисунок1), которая также преобразуется в график.

Таблица 1 – Слова, имеющие высокую семантическую близость со словом «excel»

Слово	Мера семантической близости (от 0 до 1)	Кол-во упоминаний в изучаемой выборке
exel	0,8737	34
exsel	0,7910	23
excell	0,7285	13
exell	0,7036	12
ecxel	0,6900	2

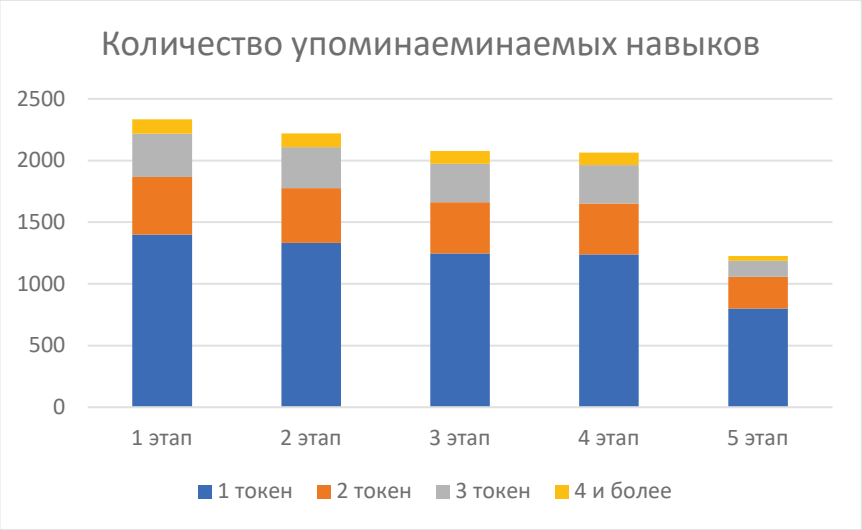

Рисунок 1 – Количество полученных навыков при использовании стандартного конвейера

Следующий шаг в задаче извлечения признаков – настройка этапов и порогов процесса обработки ССТД, который приведёт к формированию потенциальных признаков. Разделение текста из поля «профессиональные навыки» на потенциальные признаки в данном случае происходит путем использования разделителей, которые присутствовали в исходных данных (запятые или теги, в зависимости от времени создания резюме на портале). Для других данных может использоваться выделение n-gram различной длины, исходя из их встречаемости в тексте. Для обработки текстов был разработан прототип приложения, серверная часть которого написана на языке Python с использованием ряда библиотек (Pandas, Gensim, NLTK, MyStem, Plotly и т.д.). На данной стадии разработки исходный код приложения не опубликован. С помощью него готовятся данные, которые станут основой для создания визуальной модели. Исходный признак, итоговый (представляющий группу признаков) и этапы трансформации, которые были произведены, представляются в виде таблиц. Пример трансформации данных в таблице 2.

Таблица 2 – Примеры трансформации данных

Признак до обработки	Предобработка	Обработка сокращений и ошибок	Объединение в группы
1c	1c	1c	1c
1C	1c	1c	1c
работа в программе 1c предприятие	работа в программе 1c предприятие	работа в программе 1c предприятие	1c предприятие
1C предприятие 1c бухгалт	1c предприятие 1c бухгалт	1c предприятие 1c бухгалтерия	1c предприятие 1c бухгалтерия

Далее, эти данные используются для рендеринга интерактивной визуальной модели в браузере, используя инструменты, написанные на языке JavaScript. Пример интерактивной визуальной модели с использованием диаграммы Санкей, построенной на основе таблицы n-gram, связанных с упоминанием профессиональных навыков, содержащих в себе упоминание «1С», представлен на рисунке 2.

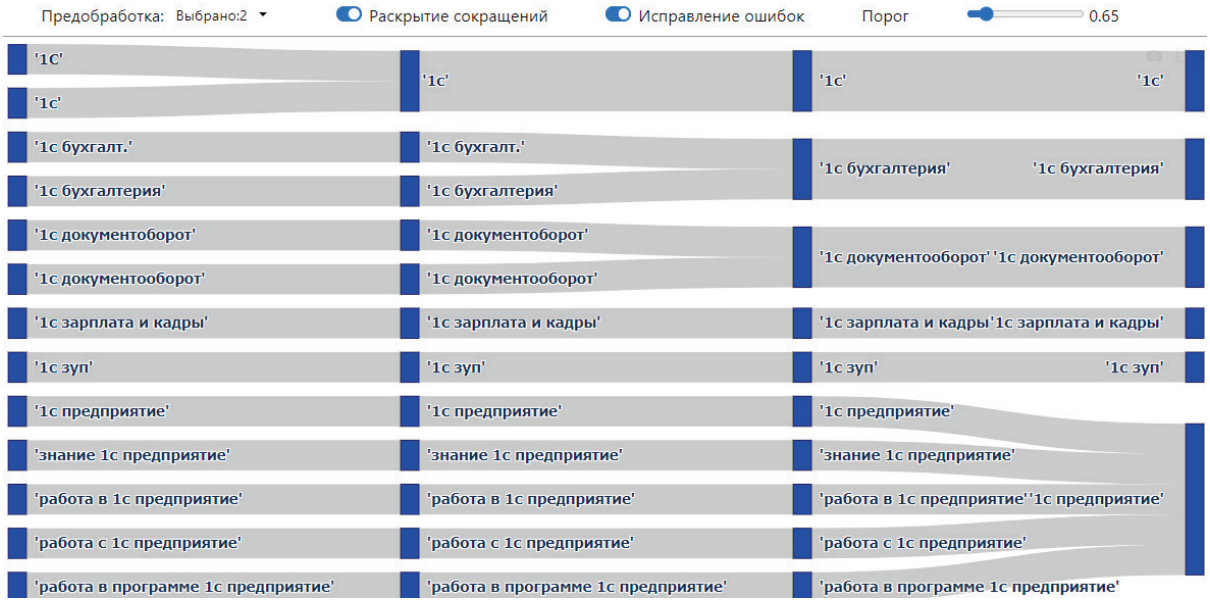

Рисунок 2 – Интерфейс интерактивной визуальной модели

Помимо визуального анализа, изменения настроек обработки и порога группировки для всего конвейера обработки в целом, интерактивная визуальная модель позволяет изменить формирование конечных признаков для случаев, которые являются «исключением из общего порога», разъединив группу и сформировав их неё новую, или объединив несколько групп в одну (рисунок 3). Предполагается, что порог должен быть выбран исследователем таким образом, чтобы для качественного извлечения признаков понадобилось как можно меньше ручных вмешательств.

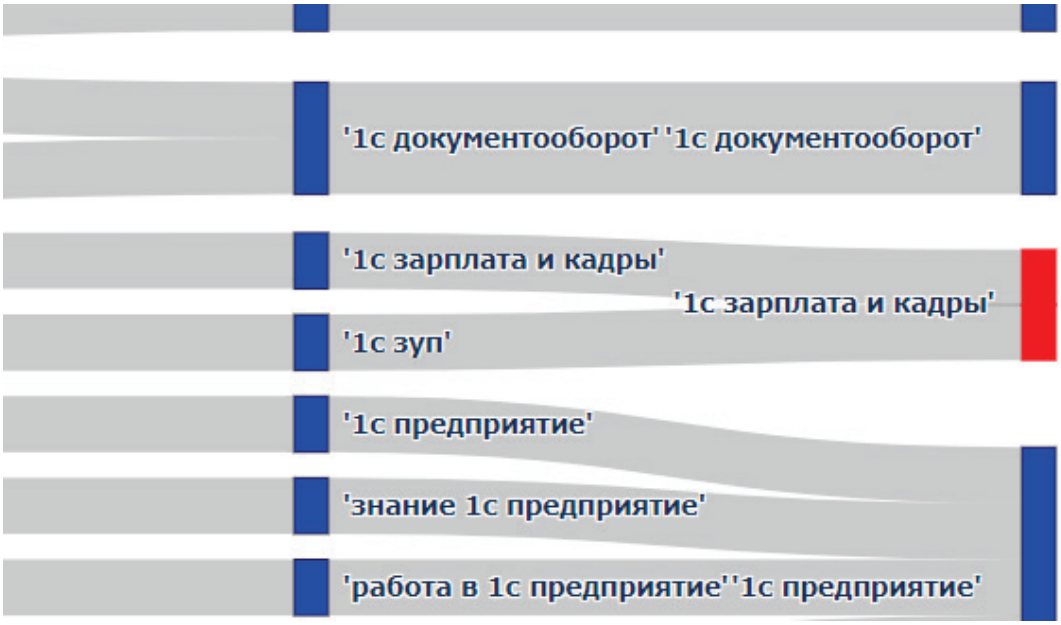

Рисунок 3 – Фрагмент интерактивной визуальной модели после ручного объединения «1с зарплата и кадры» и «1с зуп»

Также, используя интерактивную визуальную модель, можно сфокусироваться на пути формирования одного признака. Пример представлен на рисунке 4.

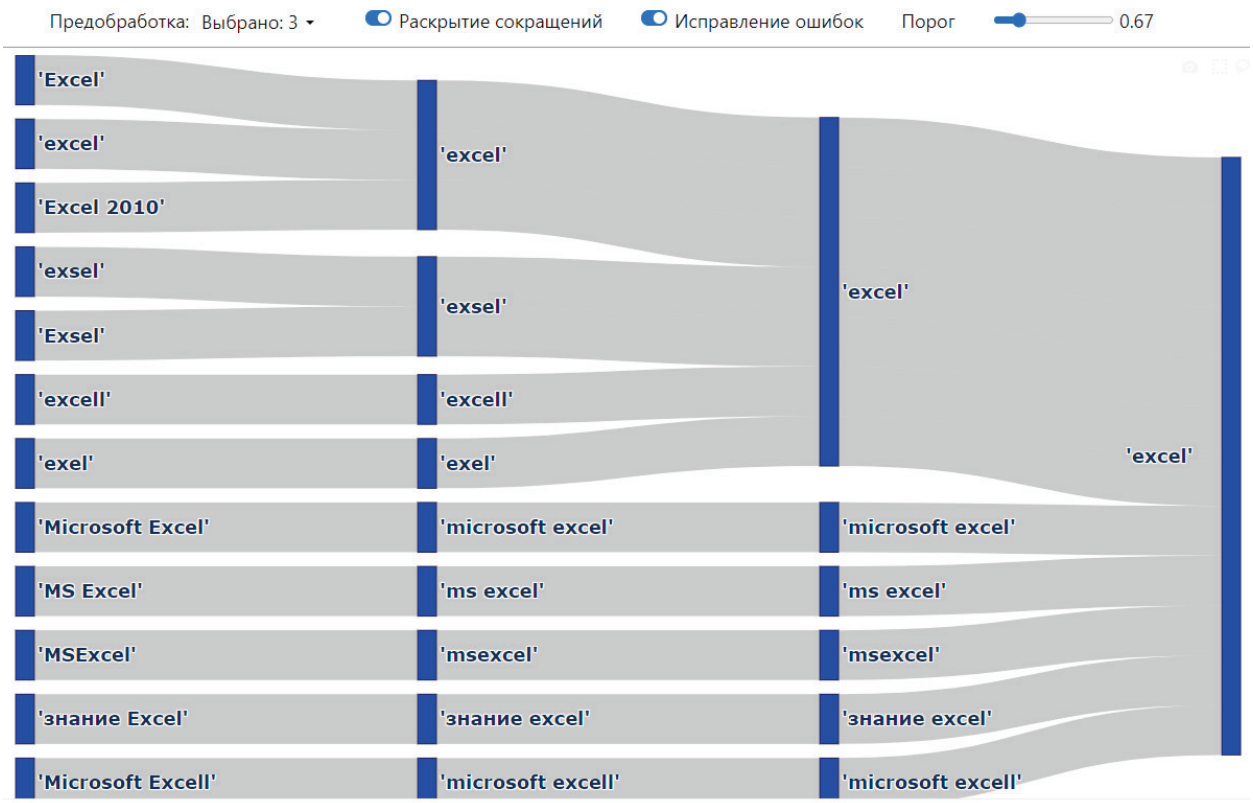

Рисунок 4 – Интерфейс визуальной модели процесса формирования одного признака

Диаграммы Санкей реализованы в ряде программных средств, поддерживающих визуализацию данных. Преимуществом разработанного программного обеспечения, поддерживающего предложенную визуальную модель, перед аналогами являются: адаптация диаграммы под визуализацию процесса трансформации n-gram и возможность ручного редактирования диаграммы с сохранением параметров для дальнейшей обработки данных.

Использование визуальной модели повышает наглядность и прозрачность использования модели интеллектуальной обработки ССТД и ускоряет процесс создания исключений в правилах группировки признаков по порогу семантической близости. Интерактивная визуальная модель позволяет настраивать параметры формирования набора признаков без изменения исходного кода обработки ССТД, благодаря чему с ней может работать эксперт в исследуемой предметной области, не имеющий навыков обработки данных с использованием языков программирования. Всё это приводит к уменьшению трудоемкости процесса, и повышению качества обработки данных за счет уменьшения количества потенциальных ошибок в процессе обработки ССТД.

4. Заключение

Задача настройки обработки ССТД является сложной и многофакторной, зависит от особенностей текста и решаемой задачи. В модели интеллектуальной обработки ССТД обычно присутствуют типовые этапы, из которых аналитик должен собрать конвейер для обработки имеющегося набора данных. Это трудоемкая задача, которую можно автоматизировать, но полностью автоматическое решение не будет учитывать особенности предметной области и, соответственно, не сможет обеспечить построение качественного конвейера. Для построения качественной модели анализа, в которую будут включены признаки, извлекаемые из ССТД, необходимо привлечение эксперта в предметной области, который может не быть специалистом

в компьютерной обработке данных и имеет ограничение на объем обрабатываемой информации и невысокую скорость обработки массивов информации.

Визуализация данных – это способ предложить эксперту в предметной области инструменты визуальной аналитики, упрощающие работу с данными и ускоряющими процесс их анализа. Так как в процессе обработки ССТД результаты обработки одного этапа влияют на результаты следующего, для визуализации процесса предложено использовать визуальную модель на основе диаграммы Санкей. Использование визуальной модели в процессе настройки модели интеллектуальной обработки ССТД из резюме соискателей продемонстрировало возможность использования предложенного подхода для визуальной аналитики и корректировки процесса извлечения признаков. Это позволило повысить наглядность и прозрачность процесса обработки ССТД, а также открыло возможность привлечь к процессу экспертов в предметной области, что повысило контролируемость и качество ССТД, которые будут использоваться в моделях анализа данных.

Разработанную визуальную модель возможно применять и в других предметных областях, например, при извлечении признаков, описывающих пациентов, из данных систем здравоохранения или при анализе данных из социальных сетей. Дальнейшие исследования будут направлены на развитие подходов к комплексной визуализации ССТД и процессов их обработки с целью снижения трудоемкости и уменьшения количества ошибок при решении различных задач обработки естественного языка и доработку программного обеспечения.

5. Список источников

[1] Hickman L., Thapa St., Tay L. / Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations // Organizational Research Methods. 2020. No. 25. DOI:10.1177/1094428120971683.

[2] Secondary Analysis of Electronic Health Records / Komorowski M., Marshall D., Salciccioli J., Crutain Y. 2016. 427 p. DOI:10.1007/978-3-319-43742-2_15.

[3] McInnes L., Healy J. / UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction // ArXiv e-prints 1802.03426. 2018.

[4] Weitz D. / Slope Charts, Why & How // Towards Data Science. 2020. [Электронный ресурс]: URL: https://towardsdatascience.com/slope-charts-why-how-11c2a0bc28be (дата обращения: 21.06.2023).

[5] Riehmann P., Hanfler M., Froehlich B. / Interactive Sankey diagrams. Proceedings - IEEE Symposium on Information Visualization, INFO VIS. // 2005. P. 233-240. DOI:10.1109/INFVIS.2005.1532152.

[6] Клышинский Э.С., Ганеева В.А. / Метод визуальной интерпретации статического векторного пространства Word2Vec // GraphiCon 2022: труды 32-й Межд. конф. по компьютерной графике и машинному зрению (Рязань, 19-22 сент. 2022 г.). – М.: Институт прикладной математики им. М.В. Келдыша РАН, 2022. С. 297-303 DOI:10.20948/graphicon-2022-297-303.

[7] Alharbi M., Laramee R., / SoS TextVis: An Extended Survey of Surveys on Text Visualization // Computers. 2019. Vol.8. №17. DOI:10.3390/computers8010017.

[8] Kucher K., Kerren A. /Text visualization techniques: Taxonomy, visual survey, and community insights // 2015 IEEE Pacific Visualization Symposium (PacificVis), Hangzhou, China. 2015. P. 117-121. DOI:10.1109/PACIFICVIS.2015.7156366.

[9] Макарова Е.А., Лагерев Д.Г. / Использование визуальных моделей для разведочного анализа слабоструктурированных текстовых данных // GraphiCon 2022: труды 32-й Межд. конф. по компьютерной графике и машинному зрению (Рязань, 19-22 сент. 2022 г.). – М.: Институт прикладной математики им. М.В. Келдыша РАН, 2022. – С. 1094-1105.

[10] Макарова Е.А., Лагерев Д.Г. / Модель обработки слабоструктурированных текстовых данных на русском языке для интеллектуальной поддержки информационного управления в динамических организационных системах // Модели, системы, сети в экономике, технике, природе и обществе, 2022. № 3. С. 104-125.

[11] Макарова Е.А. / Обработка слабоструктурированных текстовых данных для использования в моделях анализа // Информационные и математические технологии в науке и управлении, 2023. № 1(29). С. 178-189.
[12] Паклин Н.Б., Афанасьев В.В. / Оптимальное квантование для повышения качества бинарных классификаторов // Искусственный интеллект. 2013. Вып. 4. С. 392-399.
[13] «Работа в России»: обработанные и объединенные сведения о вакансиях, резюме, откликах и приглашениях портала trudvsem.ru // Роструд; обработка: Бабушкина В.О., Тимошенко А.Ш., Инфраструктура научно-исследовательских данных, АНО «ЦПУР», 2021. Доступ: Лицензия CC BY-SA. [Электронный ресурс]: URL: http://data-in.ru/data-catalog/datasets/186/. (дата обращения: 22.04.2022).